LINDAT Results on various metrics

Below are the LINDAT en-hi details and LaBSE sentence embedding Cosine similarity scores are computed on the LINDAT English-Hindi of total 2,73,885 sentence pairs.

LINDAT en-hi filters

After filters	No. of Sentences retained	No. of sentences pruned
Filter1 (len(5-100))	1,71,912	1,01,973
Filter2 (Remove en in hi text and non-ascii values in en-hi)	1,52,053	19,859

LaBSE Cosine similarity (en-hi)

No. of sentences	1,52,053
Maximum score	0.990
Minimum score	- 0.023
Average score	0.763

1. **The below were computed on whole 1,52,053 sentences pairs each.**
2. **<u>hi & te:</u>** ***Sentences from filter2 (i.e En-Hi & En-Te)***
3. **<u>hi^ & en^:</u>** ***Translated "English sentences to Hindi" & "Hindi sentences to English" respectively using Vandan's MT"(Using Lindat En-Hi parallel corpus).***

Scores with Bleu, Chrf++

Metrics		hi^-hi	en^-en (hi)
Bleu	Min_bleu	0.0	0.0
	Max_bleu	1.0	1.0
	Avg_bleu	0.1644	0.1545
Chrf++	Min_chrf++	0.0	0.0
	Max_chrf++	1.0	1.0
	Avg_chrf++	0.3860	0.4105

LINDAT WER score

	hi^-hi	en^-en
Min_WER	0.0	0.0
Max_WER	0.91	0.75
Avg_WER	0.0098	0.0087

Filter 3

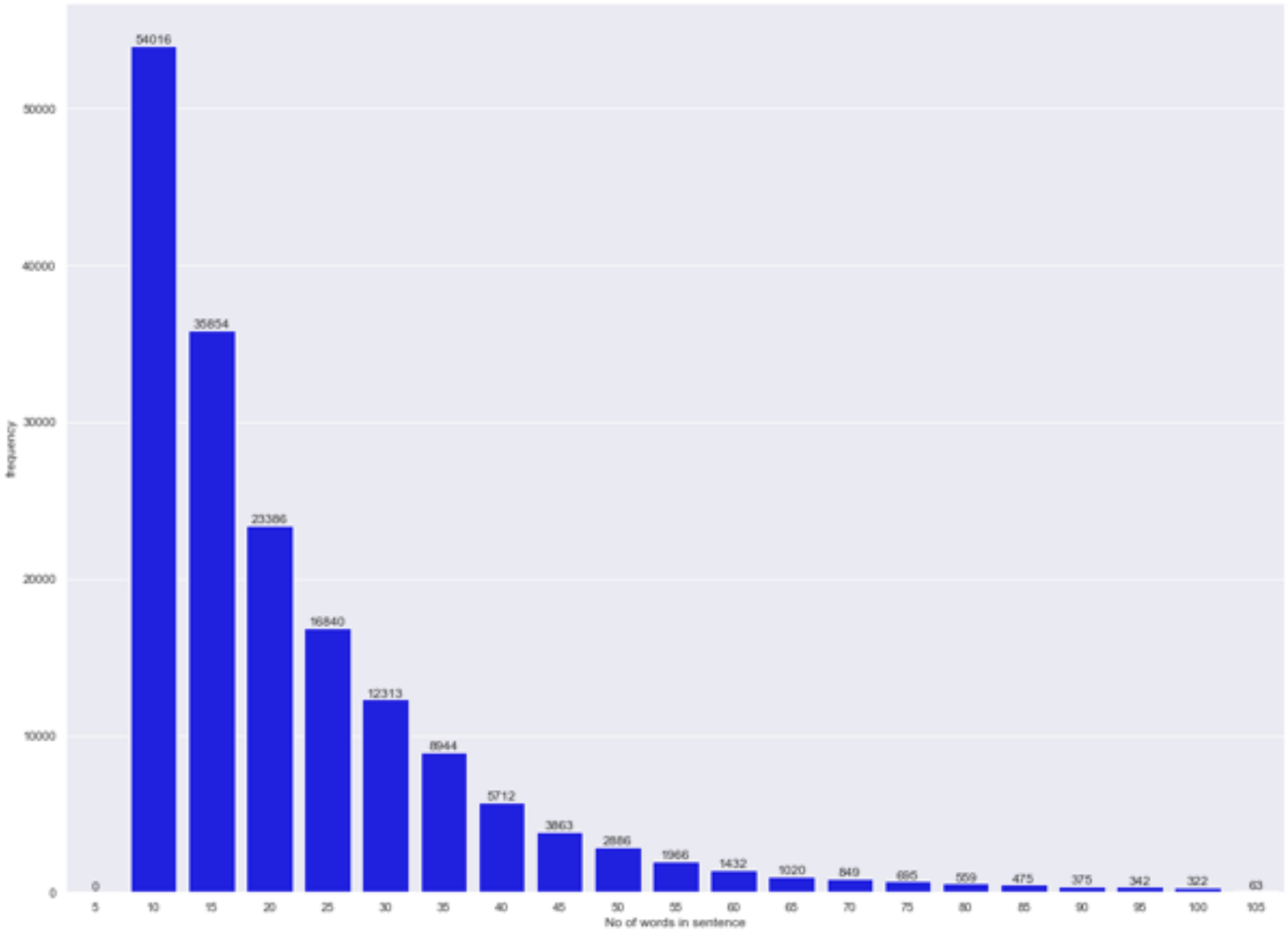
Corpus > Avg_score	Number of sentences
En-Hi > Avg_{bleu, Chrf++}	24,956
En-Hi (24,956 sentences) > Avg_{LaBSE}	Clean : 22,625
	Pruned : 2231

<u>Key points :</u>

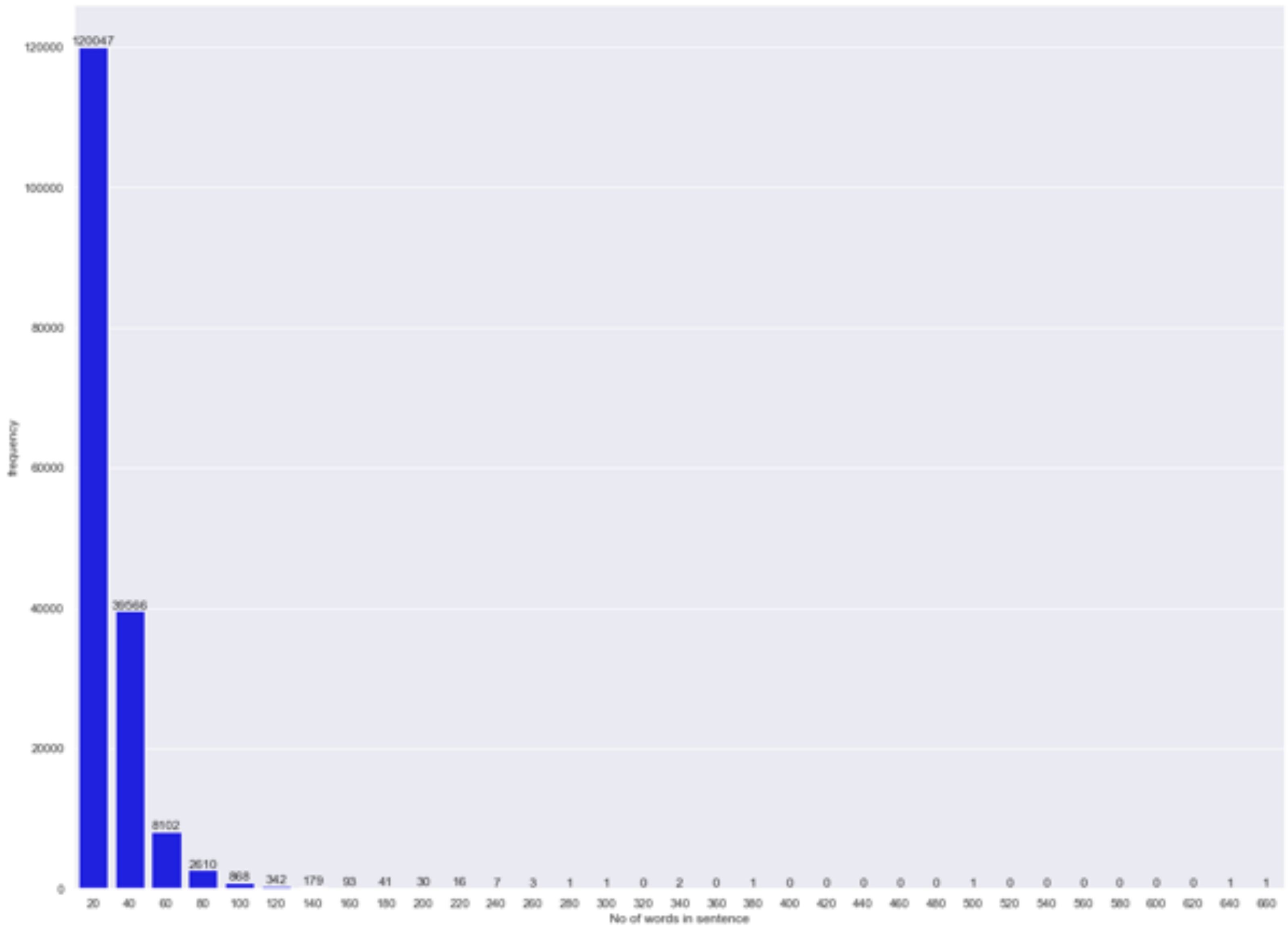
1. Couple of main points :

  1. **Time taken to translate whole 1,52,053 sentences from english to Hindi and Hindi to English - 41.2 hours each**
  2. **Calculation of WER, Bleu scores : 10 minutes for each.**
  3. **Calculation of Chrf++ scores : 17,000 sentences in 2.5 hours.**

Sentence distribution Plots after filter 1 :

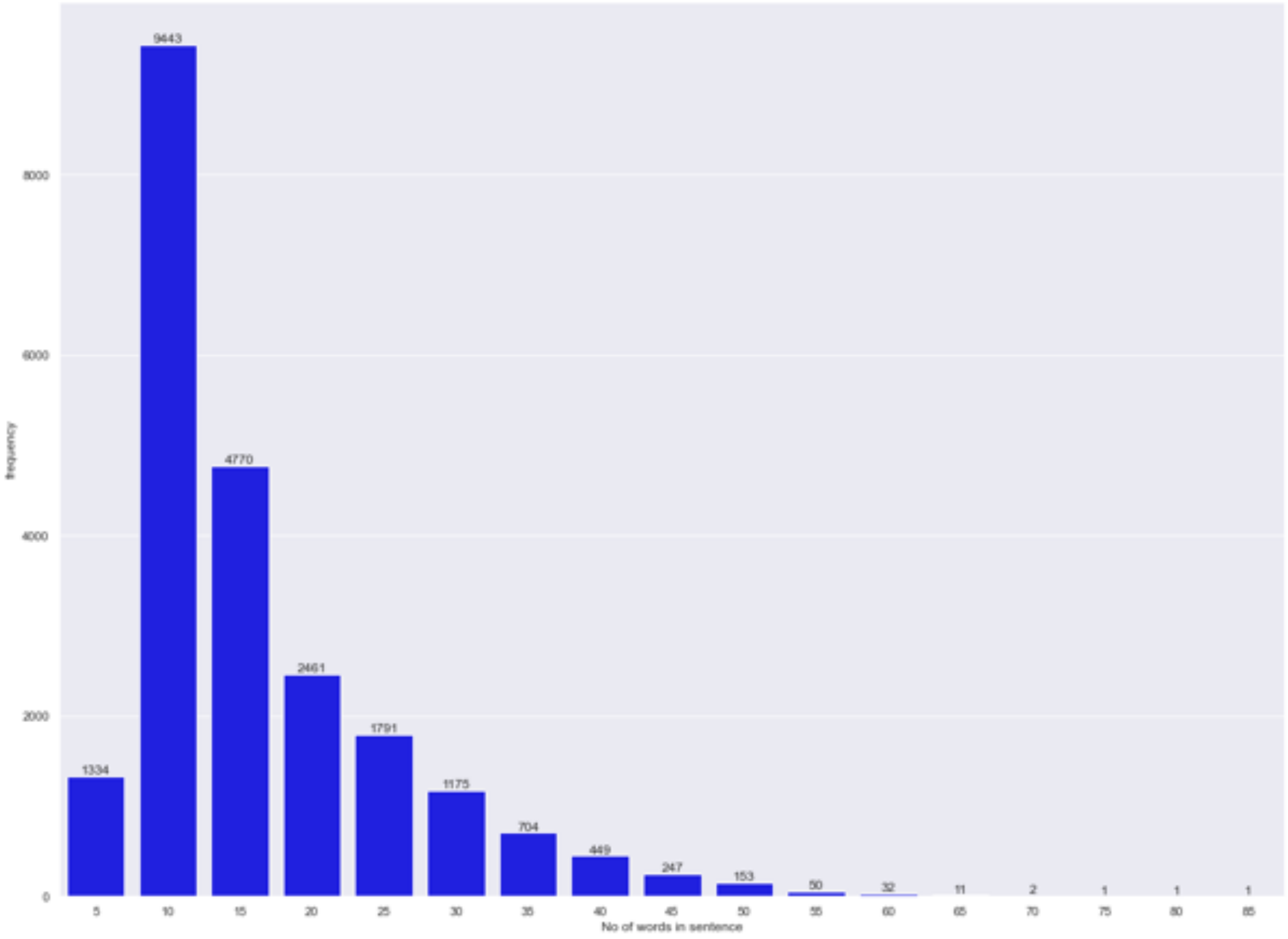


Sentence distribution of LINDAT Hindi text after filter1

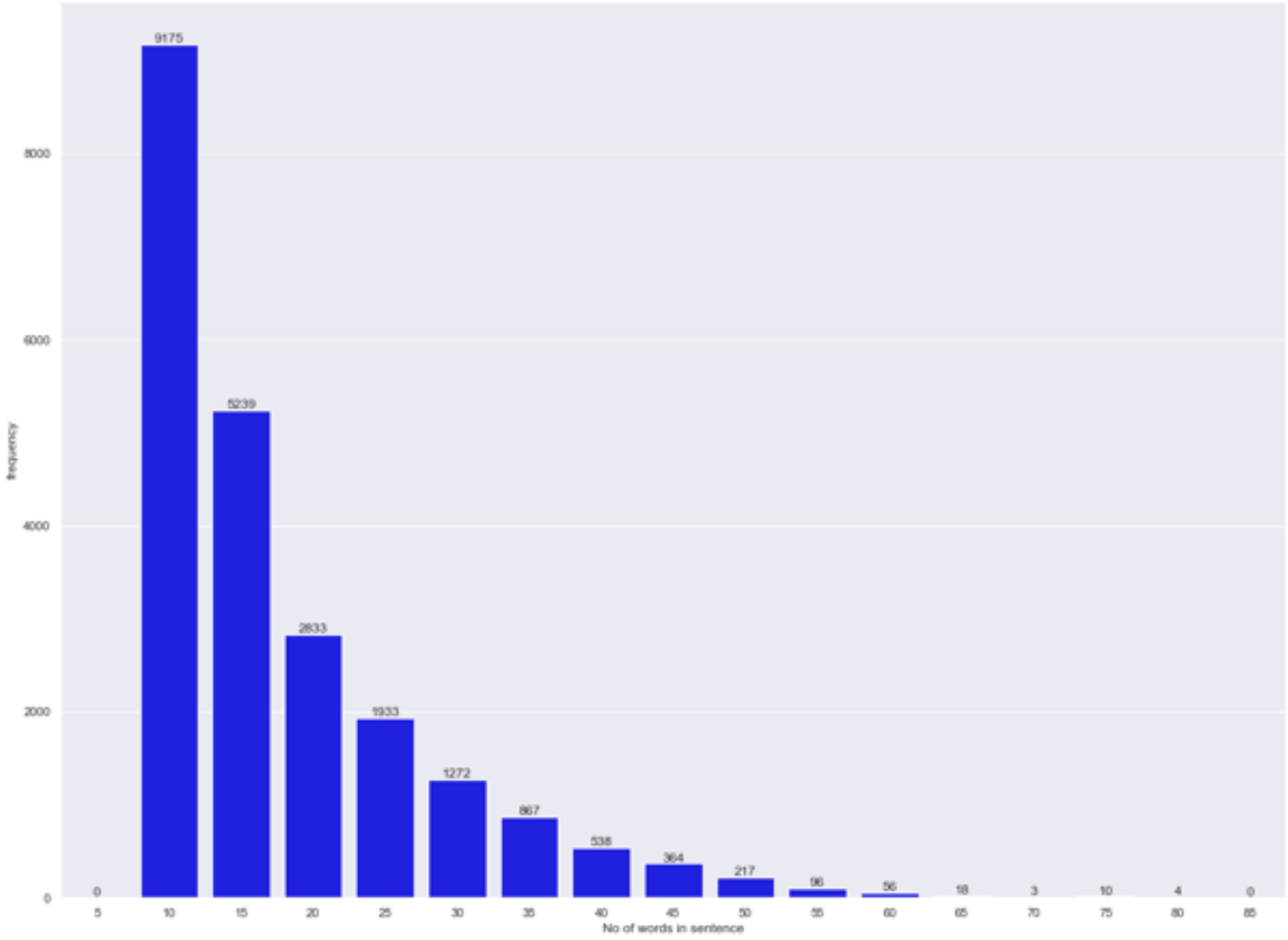


Sentence distribution of LINDAT English after filter1

Sentence distribution Plots after filter 3 :



Sentence distribution of LINDAT English after filter3



Sentence distribution of LINDAT Hindi after filter3